R E P O R T R E S U M E S

ED 012 922 AL 000 638

ENGLISH LANGUAGE PROFICIENCY TESTING AND THE INDIVIDUAL.

BY- HOLTZMAN, PAUL D.
PENNSYLVANIA STATE UNIV., UNIVERSITY PARK

PUB DATE 27 APR 67

EDRS PRICE MF-$0.25 HC-$0.60 15P.

DESCRIPTORS- *LANGUAGE TESTS, *FOREIGN STUDENTS, TESOL, TEST VALIDITY, TEST RESULTS, TEST INTERPRETATION, STUDENT TESTING, TESTING PROBLEMS, LANGUAGE ABILITY, SECOND LANGUAGE LEARNING, *DATA ANALYSIS, *FACTOR ANALYSIS,

THE AUTHOR POINTS OUT PROBLEMS IN TEST RESEARCH AND INTERPRETATION, SOME OF WHICH ARE DUE TO CONFLICTS BETWEEN THE FINDINGS OF THE DATA ANALYST WHO IS RESTRICTED TO BASING HIS DECISIONS ON SELECTED DATA ONLY, AND THE TEST INTERPRETER WHO IS AWARE OF VARIABLE VALIDITIES OF SUCH UNTESTED FACTORS AS SITUATIONAL ANXIETY, PERSONALITY, MOTHER-TONGUE INFLUENCES, CULTURAL CLASH, AND SENSE OF COMMUNICATION. HOWEVER, THE AUTHOR FEELS IN SPITE OF THESE AND OTHER SHORTCOMINGS, THERE ARE A NUMBER OF REASONS FOR CONTINUING TO DO FACTOR ANALYSIS OF TEST RESULTS. ONE FACTOR, "FEEDFORWARD," BASED ON THE PSYCHOLOGY OF PERCEPTUAL EXPECTANCE, DEALS WITH SETS OF THE CATEGORIES THAT INDIVIDUALS HAVE AVAILABLE FOR THE PROCESSING OF ANY INTERNAL AND EXTERNAL PERCEPTIONS INCLUDING THOSE FOR LANGUAGE RECEPTION AND PRODUCTION. A VALID TEST OF LANGUAGE PROFICIENCY WOULD BE A TEST OF THE CATEGORIES THAT THE SUBJECT BRINGS TO ANY PROCESSING OF THE LANGUAGE. THE AUTHOR REVIEWS RECENT AND CURRENT RESEARCH WHICH IS CONCERNED WITH THE FACTOR OF "REDUNDANCY UTILIZATION", THE ABILITY OF THE NATIVE SPEAKER TO PREDICT SEQUENTIAL LANGUAGE SIGNALS AS CONTRASTED WITH THE NON-NATIVE SPEAKER'S DEPENDENCY ON INTERPRETING EACH WORD ON THE BASIS OF THE SIGNAL ITSELF. THIS WORKPAPER WAS PRESENTED AT THE ATESL SEMINAR IN AUSTIN, TEXAS, APRIL 27, 1967. (AM)

ED012922

ENGLISH LANGUAGE PROFICIENCY TESTING AND THE INDIVIDUAL

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

A workpaper for the ATESL Seminar
on testing

Houston, Texas
April 27, 1967

by Paul D. Holtzman

"PERMISSION TO REPRODUCE THIS [redacted] MATERIAL HAS BEEN GRANTED
BY Paul D. Holtzman

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE [redacted] OWNER."

THE PENNSYLVANIA STATE UNIVERSITY
Graduate School Language Testing Center
for International Students
University Park, Pennsylvania

AL 000 638

We might parody thus: A test interpreter looking at a set of ELP scores exclaimed that Mr. Kashimura will do very well in his petroleum engineering curriculum. His assistant asked, "How can you tell?" "Because he is obviously highly motivated."

The Educational Testing Service, as a corporate data analyst, carefully avoids statements of criteria levels. The Pennsylvania Department of Public Instruction has contracted with ETS to test candidates for certification as secondary school language teachers. As corporate test interpreter, the DPI established a criterion score for certification. It was inevitable that a student would come along with a record of long and--by other criteria, successful--study in a language and with an exceptionally effective practice teaching experience who scored just one point below the required level on the ETS-MLA test. Test interpreters in our College of Education are up in arms. They perceive the other data but, by law, the DPI Bureau of Certification cannot.

B. The data analyst--so long as he is only the data analyst--is never aware of the human consequences of his decisions. Any sensitive test interpreter cannot escape this knowledge.

In the case just cited (and it is a real case) the test interpreters are upset by a host of human consequences: effects upon the would-be teacher and a real loss to students somewhere exposed to a less effective teacher who may have scored just above the criterion.

Confrontation of human consequences of test score interpretations is an almost daily occurance in my office. Mr. Kuo was one of five graduate students in Chemistry who were told that they must achieve minimum English proficiency by the end of their first term. If not, their assistantships would not be renewed the following year. End-term test results showed that the other four had achieved "minimum proficiency" or better, that Mr. Kuo had not, and that Mr. Kuo had made more progress than the

other four. But to him the others had succeeded where he had failed and this made it almost impossible for him to face his Department.

Mr. Bolivar was a perennial student of English. After two years he had completed all of the requirements for a master's degree in petroleum engineering but had not achieved the "minimum proficiency" required of all candidates by the Graduate School--at least not according to the test scores. When, on the basis of human consequences, it was reported that he had at last met the requirement, even the foreign student adviser was critical--of the ELP test.

On the other hand, there was Mr. Pak. He completed his courses, left the University, and submitted a thesis from somewhere in the world. For several terms his name was removed from the graduation list because there was no evidence of his meeting the ELP requirement. At last he took the TOEFL in Japan. For several days past the printer's deadline the Graduate School held the graduation list. At last the TOEFL score arrived: about 330. The ensuing conversations between Mr. Pak's Department and the Dean of the Graduate School no doubt dealt with human consequences The degree was awarded.

I am sure that all members of this seminar can match me story for story except, possibly, those who meet my definition of "data analyst."

C. The data analyst, given adequate reliability, bases his decisions on assumed validity. The test interpreter is aware of variable validities attributable to untested factors such as situational anxiety, personality, mother-tongue influences, cultural clash, and sense of communication.

Consider, for instance, the statistical process of item analysis. Under the heading of "discrimination" what do we look for? The extent to which each item discriminates between those who do well and those who do poorly on the total test. Thus, in essence, we seek to improve reliability, assuming validity for total test scores. (see also Problem 3)

Teachers of foreign languages in our schools and colleges have long been aware of the fact that their students who may be academically equal, according to tests, will vary over a wide range of abilities in using the foreign language to communicate with natives.[2] Yet it is this ability to communicate in the language, rather than knowledge of the language, that we are trying to test. As far as I know there is no dependable test of this ability for second language learners. We are just beginning to develop one for American students.[3]

D. The data analyst is engaged in <u>transmission</u> of data; the test interpreter is engaged in human interaction.

In the transmission of data, meaning is irrelevant so long as the data received are the same as the data sent. In a communication transaction, meanings are more important than the data. Test interpretation is, obviously, a communication transaction. The test interpreter, then, must take into account the meanings of what he has to report to the student, to his adviser, to his department.

<u>Problem 2</u>. The goal of ELP test research seems to be one of devising means of validating <u>all</u> ELP decisions on the basis of data analysis. Yet validation is <u>always</u> statistical. This means that we can only provide valid descriptions of <u>groups</u> and can <u>never</u> account for all variances from derived statistical norms.

We find that a test has high reliability but never perfect reliability. What does this mean? It means that for some individuals the test lacks consistency. We validate reliable tests by determining correlations between test scores and those derived from some criterion measure (hopefully also reliable). The correlation is

---

[2]See for instance Kenneth L. Pike, "Nucleation," <u>The Modern Language Journal</u>, November, 1960, 291-295. Reprinted in Harold B. Allen, <u>Teaching English as a Second Language</u>, McGraw-Hill, 1965, pp. 67-74.

[3]Called a "Commsense Inventory," it attempts to test the extent to which students are audience-centered in their concepts of speaking and writing.

never perfect. Darrell Huff reminds us to "keep in mind that a correlation may be real and based on real cause and effect--and still be almost worthless in determining action in any single case."[4]

Problem 3, Lines of research are subject to the common criticism of any detective work: "They look everywhere until they find a suspect, but they're likely to concentrate on him from then on."[5] To throw in another McLuhan aphorism, "As we begin, so shall we go."[6]

Again, our statistical methods help us (or force us) to concentrate on "a suspect" or test variable. One already mentioned is item analysis which we use to increase not the validity but the reliability of our tests, to improve "concentration" on the test variable.

A popular statistical method that does not inherently force such concentration but may be used to do so is factor analysis. Having arrived at the conclusion that we had isolated a general ELP factor,[7] we have eliminated tests which did not "load" on that factor. Yet the outcome of a factor analysis is determined by what it put in From our own research, here are some examples which have not been previously reported

First, Table I shows the kind of concentration that accrued when we proceeded from the results reported earlier[8] to refined tests and a new factor analysis. These are data derived on the Penn State ELP test from Indiana University students.

---

[4]How to Lie with Statistics, W. W. Norton, 1954, p. 93.

[5]From Harry Kemelman, Friday the Rabbi Slept Late, Fawcett Crest, 1965, p. 114.

[6]The Medium Is the Massage, p. 45.

[7]See Richard E. Spencer and Paul D. Holtzman, "It's Composition--But Is It Reliable?" College Composition and Communication, May, 1965, pp. 117-121.

[8]Ibid.

TABLE I

FACTOR ANALYSIS SHOWING ONLY
HIGHLY SIGNIFICANT LOADINGS

Test	I	II	III	IV	V
1. Sound discrimination	.51			.46	.48
2. Accuracy of dictation					.97
3. Written structure			.86		
4. Attitude toward English			.73		
5. Word fluency	.76				
6. Paragraph reading	.82		.42		
7. Scrambled text				.92	
8. Vocabulary		.60	.53		.43
9. Rated intelligibility		.95			
10. Rated listening ability	.87				
11. Oral stress		.95			
% of variance accounted for	40.5	14.2	12.4	10.9	9.6

Within the process of factor analysis is another source of perception control: the labeling of factors. In earlier studies we found a general ELP factor plus others which we labeled "academic ability or intelligence" and "attitude toward English." The reader might try his hand at labeling the five factors above. It is a dangerous practice.

What should be noted in Table I is that, whatever the factors, they seem to account for a large portion of the variance and they seem to be getting at significant aspects of English language proficiency.

By contrast, however, Table II shows what happens if we introduce more test variables, all presumably designed to get at the same abilities assessed by the Penn State ELP subtests. These data are from the same subjects (Indiana University) and include the data which produced Table I.

TABLE II

FACTOR ANALYSIS SHOWING ONLY
HIGHLY SIGNIFICANT LOADINGS

Test	I	II	III	IV	V
1. Michigan test of aural comprehension	.45			-.71	
2. Michigan test (total)	.44	-.53		-.62	
3. TOEFL listening comprehension	.99				
4. TOEFL structure	1.05				
5. TOEFL vocabulary			-.60		
6. TOEFL reading comprehension		-.73			
7. TOEFL writing ability	.47		-.54	.73	
8. TOEFL total	.47		-.40		
9. Listening and speaking*		-.63			
10. Initiation and conversation*		-.72			
11. Interest and motivation*		-.73			
12. Performance-phonology*			-.66		
13. Performance-structure*	-.52	-.44	-.44		
14. Aural comprehension*					.99
15. Initiation and conversation**			-.89		
16. Interest and motivation**	.76				
17. Performance-writing**			-.66		
18. Performance-longer writing**					.54
19. PSU sound discrimination			-.44	-.52	.49
20. PSU accuracy of dictation	.43		-.49		
21. PSU written structure		-.67			
22. PSU attitude toward English		-.43			
23. PSU word fluency			-.47	.43	
24. PSU paragraph reading				-.75	
25. PSU scrambled text					.58
26. PSU vocabulary					
27. PSU rated intelligibility			.71		
28. PSU rated listening ability					.42
29. PSU oral stress					.56
% of variance accounted for	6.21	4.61	3.49	2.93	2.77

*instructor ratings--classes in spoken English
**instructor ratings--classes in written English

Table II is presented for no purpose other than to confirm the idea that factor analysis results are a function of the data submitted or, as someone has put it more succinctly: "Garbage in--garbage out." Little of the variance is accounted for. Basically similar tests load on different factors. What this tells us is that we are dealing with such complexities of interacting variables as to challenge reduction to simple scores and assessment on the basis of those scores alone.

Further, it seems clear that concentration on one variable or set of variables to the exclusion of others in test development and application is fraught with dangers.

II. Perceptual expectancy and the search for an integrative factor

In spite of the results cited above, there are a number of reasons for continuing to do factor analysis of test results. Some of these have to do with refinement of the tests themselves; some have to do with diagnosis on the basis of more or less independent factors; some have to do with teaching. If two skills are closely related, if they are co-incident, must each be taught? The teacher is always looking for integrative factors--for the skills whose development automatically increase abilities in other skills. In our program, for instance, we have found that emphasis on the learning of unstressing patterns and rhythms obviate the necessity of drilling on certain phonemes. The unique unstressing behaviors of English speakers are, we believe, an integrative factor.

A far more important integrative factor--of significance in both testing and teaching--seems to be ignored except in a few recent research efforts. This is what I. A. Richards might call a feedforward factor. It is based in the psychology of perceptual expectancy. It deals with sets of the categories that individuals have available for the processing of any internal and external perceptions including those for language reception and production.

"It's all Greek to me" is a statement that anything that does not fit in the category "English" is perceived in the category "Greek;" or that any English that I don't understand might as well be filed with Greek which I also don't understand. When a speaker of the General American dialect visits parts of Texas he finds that his expectancy for the first person singular pronoun is often violated. Since he cannot "file" it under /aI/ and he does not have the category /a:/, he "files" it under /a/. He is helped in this, of course, by the fiction writers who spell it, Ah.

Learning a new language is a process of structuring new, complex sets of perceptual expectancies--both for reception and for production of the language. It might follow, then, that a valid test of language proficiency would be a test of the categories that the subject brings to any processing of the language. These categories necessarily include expectancies for not only words or vocabulary but for all of the variables that we have been trying to test. They would seem to comprise an integrative language factor.

Most of our tests, however, force the student to respond with the examiner's categories rather than his own; with the test-writer's words and sounds and structures whether or not they are also the student's. The multiple choice item is a case in point. The language of response is chosen from the limited categories drawn from the perceptual possibilities projected by the test constructer.

Can we assess the appropriateness of a language-learner's categories for language perception? The question can't be answered, of course. But several lines of research suggest that the answer might be yes, within limits.

All of the research reviewed below is concerned, in one way or another, with a factor of "redundancy utilization."[9] In normal (first language) function we know what to expect--and therefore what categories to have available--on the basis of redundancy in the language, in the situation, in the "image" of the speaker or listener or writer or reader, in the context, and so on. We know what to expect. We use the language signal as best we can to confirm or deny the expectancy.

Alan C. Nichols has found that American and foreign students have different patterns of error in writing dictated sentences categorized has having "high naturalness." For successive words, native speakers made increasing errors toward the middle of each sentence and fewer errors toward the end. The foreign students

---

[9]From Wendell W. Weaver and Albert J. Kingston, "A factor analysis of the Cloze procedure and other measures of reading and language ability," Journal of Communication, XIII:4, 1963, pp. 252-261.

(Japanese) made errors at about the same rate in all parts of each sentence.[10] It seems appropriate to conclude that the native speakers were able to make use of the redundancy of the "highly natural" sentences to predict with increasing accuracy what would follow. The Japanese students were less able to predict and more dependent on interpreting each word on the basis of the signal itself.

In later research, Nichols administered his test of Memory Span for Immediate Recall (MSIR) along with the Penn State tests. MSIR scores correlated most highly with those subtests which might seem to require some redundancy utilization and with all but one of the subtests that require the subject to use his own language.

TABLE III

RANK ORDER CORRELATIONS
BETWEEN MSIR AND PENN STATE SUBTESTS

Rated Intelligibility	.64
Rated Listening Ability	.64
Dictation	.58
Sentence Completion	.57

I have taken the term "redundancy utilization" from the work of Weaver and Kingston[11] who compared the MLAT, a battery of comprehension and vocabulary tests, and Cloze procedure in a factor analysis. Having put Cloze procedure in, they got Cloze procedure out as a factor and then concluded that it was perhaps little more than "an interesting curiosity." Carroll and others[12] conducted a "pilot investigation" of the applicability of Cloze procedure in testing foreign language achievement.

---

[10] "Apparent factors leading to errors in audition made by foreign students," Speech Monographs, XXXI:1, March, 1964, pp. 85-91.

[11] op. cit

[12] John B. Carroll, Aaron S. Carton, and Claudia P. Wilds, An Investigation of "Cloze" Items in the Measurement of Achievement in Foreign Languages, Laboratory for Research in Instruction, Graduate School of Education, Harvard University, 1959.

Some of their conclusions include:

1. The fact that cloze scores are so highly correlated with various factors of cognitive ability when the testing is in the subject's native language raises grave question as to the potential efficacy of the cloze procedure as a measure of the subject's achievement in a foreign language. (p. 66)

2. . . . results strongly suggest that the cloze tests are in fact measuring some important facet of foreign language proficiency--but this is much more true for <u>groups</u> than for <u>individuals</u>. That is to say, if we use group results to cancel out individual variations on all the extraneous factors which may contribute to the determination of cloze test scores, the group means reflect real differences in foreign language competence. (p. 85)

3. Another kind of evidence of validity is to be found in the correlation of cloze test scores with teachers' grades. It is a common myth among educational psychologists that teachers' grades are notoriously unreliable; but this does <u>not</u> seem to be true, necessarily, of teachers' grades in foreign language courses, which are frequently found to correlate highly enough with other variables to suggest that they are quite reliable. (p. 85)

Cloze procedure, in essence, tests for what word a reader would expect to read (or hear) where the original word has been deleted in a paragraph. Rankin, who critically evaluated Cloze procedure, [13] found highly stable correlations between scores based upon production of the original word in each case and scores based upon production of a word which "made sense" in each case. That was with native speakers of English. In the study cited above, Carroll and others compared original-word and "community of response" scoring, finding the latter slightly more reliable but slightly less correlated with CEEB scores. Last month we scored our Cloze test two ways--on the basis of original word and on the basis of "makes sense"--and found no correlation for the foreign students tested.

In a study of "Cloze procedure as a test of English language proficiency," Hopf and Spielmann found some variations attributable to the form classes of words

---

[13] E. F. Rankin, Jr., "An Evaluation of the Cloze Procedure as a Technique for Measuring Reading Comprehension, unpublished dissertation, U. of Michigan, 1957.

which happened to be deleted in two forms. This suggests that words for deletion should be selected not by chance but for the testing purpose. In a new study, Spielmann is attempting to find ways to reduce variance attributable to outside influences. He will include a hypothesis of greater validity of Cloze procedure scores when corrected for each student's Cloze ability in his native language (beginning with Spanish speakers).

Spolsky is engaged in essentially the same line of research into redundancy utilization, it seems to me, in the studies reported last year in his "Progress report, January-September 1966."[14] He hypothesizes that "overall proficiency in a language . . . may be measured by testing a subject's ability to send and receive messages under varying conditions of distortion of the conducting medium." Where our work with Cloze has begun with "distortion" of the written language, Spolsky has begun with distortion of the spoken language (by introducing white noise at discrete intensity levels).

Both Spolsky and Spielmann plan to apply their experimental tests to languages other than English. They may not have the same comparisons in mind but, in any case, should produce some data that may offer encouragement to those of us who would assess that integrative factor of linguistic perceptual expectancy (operationally defined, perhaps, as redundancy utilization).

---

[14]Bernard Spolsky, Preliminary Studies in the Development of Technqiues for Testing Overall Second Language Proficiency, Indiana University, 1966. (mimeograph)